分类号 密级

UDC注1



**硕 士 学 位 论 文**

**基于预训练视觉基础模型微调的**

**遥感图像高效分割方法**

（题名和副题名）

**赵之心**

（作者姓名）

**指导教师姓名 肖亮** 教授

**学 位 类 别 工学硕士**

**学 科 名 称 计算机科学与技术**

**研 究 方 向 深度学习**

**论文提交时间 2026年4月**

注1：注明《国际十进分类法UDC》的类号。

**声 明**

本学位论文是我在导师的指导下取得的研究成果，尽我所知，在本学位论文中，除了加以标注和致谢的部分外，不包含其他人已经发表或公布过的研究成果，也不包含我为获得任何教育机构的学位或学历而使用过的材料。与我一同工作的同事对本学位论文做出的贡献均已在论文中作了明确的说明。

研究生签名： 年 月 日

**学位论文使用授权声明**

南京理工大学有权保存本学位论文的电子和纸质文档，可以借阅或上网公布本学位论文的部分或全部内容，可以向有关部门或机构送交并授权其保存、借阅或上网公布本学位论文的部分或全部内容。对于保密论文，按保密的有关规定和程序处理。

研究生签名： 年 月 日

# 摘 要

（待修改）随着最新地理空间数据及其快速提取相关信息的方法的需求显著且日益增长，对精确地理空间信息的需求在全球范围内不断增加，以适应当前的世界需求。遥感图像分割作为基于计算机视觉的应用中的一个重要组成部分，服务于自主车辆的车道分析和无人机的地理定位等领域。近年来，深度学习方法的发展又极大地推动了遥感图像分割的发展，其在许多传统计算机视觉应用、检测和分割中显示出远超传统方法的性能。然而，深度学习方法通常依赖于大量标记数据，这在遥感图像数据中往往伴随巨量的人工成本，很难为遥感图像中的元素分配有意义的标签。而基础大模型的发展显著缓解了这些问题，其展现出的卓越特征提取和表示能力，使其在自然语言处理、计算机视觉等多个领域得以广泛应用。但想要将视觉基础模型扩展到遥感图像领域下，来实现良好的泛化性能，往往面临诸多问题。首先是遥感图像与自然图像之间的领域差距导致基础模型泛化性不佳；其次，视觉基础模型对于遥感图像中细粒度目标分割效果不稳定；除此之外，特定基础模型在遥感图像领域的分割任务拓展存在局限。为了解决上述问题，本文针对视觉基础模型在遥感图像下的微调方法展开研究，具体的工作和创新点如下：

（1）针对俯视遥感图像与近距离自然场景图像之间的领域差距问题，本文考虑设计一种基于SAM自适应微调与特征增强的遥感图像实例分割方法。通过对SAM进行自适应微调，以补充基础视觉模型在遥感领域的知识缺失，同时通过自适应多尺度融合的方式提高掩码质量，使得能够自动实现细粒度的实例分割。实验表明此方法显著提升了SAM在遥感图像实例分割场景下的精度；

（2）针对视觉基础模型对复杂背景的遥感图像中多目标、小目标的分割效果不稳定问题，本文提出了一种基于SAM辅助编码与特征混合解码的深度网络模型。旨在在保持SAM原有的泛化能力下，进一步拓展视觉基础模型的泛化性能。通过引入一种针对遥感图像更有效的解码器模块，同时通过简单的辅助编码器增强SAM在遥感图像下特征的浅层信息。实验表明提出的方法使得SAM更好地适用于遥感图像语义分割任务。

（3）针对视觉基础模型作用于遥感图像领域不同分割任务的拓展局限性问题，本文设计一种基于SAM提示微调的遥感图像交互式分割方式。通过设计使用正负提示样本的方法来激活SAM在遥感图像领域的零样本泛化能力。而对于包含语义信息的提示词，通过基础的预训练语义模型CLIP来生成语义提示嵌入，帮助整体框架更好地理解提取的全局特征，并通过不同的提示驱动对应的分割任务的实现。

**关键词**：深度学习，视觉基础模型，SAM，遥感图像分割，模型微调

# Abstract

III

In recent years, the demand for the latest geospatial data and rapid methods to extract relevant and useful information has grown significantly worldwide, driven by the increasing need for precise geospatial information to meet current global challenges. Remote sensing image segmentation, as a crucial component of computer vision applications, serves functions such as lane analysis for autonomous vehicles and geolocation for drones. With the development of deep learning methods, performance in many traditional computer vision tasks, including detection and segmentation, has greatly surpassed conventional approaches, significantly advancing remote sensing image segmentation.

However, deep learning methods typically rely on large amounts of labeled data, which entails substantial manual labeling costs in remote sensing imagery, making it difficult to assign meaningful labels to elements within such images. To address this, semi-supervised, weakly supervised, and self-supervised learning methods have emerged, aiming to reduce annotation costs and improve model performance under limited labeled data scenarios. Notably, the advancement of large models has significantly propelled the progress of self-supervised learning, demonstrating outstanding feature extraction and representation capabilities that have been widely applied across natural language processing, computer vision, and other fields.

Nevertheless, extending foundational visual models to the remote sensing image domain to achieve robust generalization faces several challenges: First, there exists a significant domain gap between overhead remote sensing images and close-range natural scene images; second, the segmentation performance of foundational visual models on complex backgrounds with multiple and small targets in remote sensing images is unstable; third, these models have limitations when extended to different segmentation tasks within the remote sensing domain.

To address these issues, this paper focuses on the generalization fine-tuning of foundational visual models for remote sensing image segmentation tasks. The specific contributions and innovations are as follows:

1. To tackle the domain gap between overhead remote sensing images and close-range natural scene images, this paper proposes a remote sensing image instance segmentation method based on SAM (Segment Anything Model) adaptive fine-tuning and feature enhancement. By adaptively fine-tuning SAM, the knowledge gap of foundational visual models in the remote sensing domain is supplemented. Meanwhile, the mask quality is improved through adaptive multi-scale fusion by redesigning and training the decoder to automatically achieve fine-grained instance segmentation. A multi-scale feature enhancement module is introduced to compensate for SAM’s lack of multi-scale features.

2. To address the unstable segmentation performance of foundational visual models on multiple and small targets within complex remote sensing image backgrounds, a deep network model based on SAM-assisted encoding and feature-mixing decoding is proposed. This aims to extend the generalization capability of foundational visual models while preserving SAM’s original generalization ability. A more effective decoder module for remote sensing images is introduced, along with a simple auxiliary encoder to enhance SAM’s shallow feature information in the remote sensing context.

3. To overcome the limitations of foundational visual models in extending to different segmentation tasks within the remote sensing domain, an interactive remote sensing image segmentation approach based on SAM prompt fine-tuning is designed. By employing positive and negative prompt samples, SAM’s zero-shot generalization ability in remote sensing images is activated. For prompts containing semantic information, semantic prompt embeddings are generated using the pre-trained semantic model CLIP, assisting the overall framework in better understanding and extracting global features, and enabling the realization of various segmentation tasks driven by different prompts.

**Keywords:**  Deep learning, Foundational visual models, SAM, Remote sensing image segmentation, Fine-tuning.

# 目 录

**[Document Title]**

[摘 要 I](#_Toc200886762)

[Abstract II](#_Toc200886763)

[目 录 IV](#_Toc200886764)

[1 绪论 1](#_Toc200886765)

[1.1 研究背景及意义 1](#_Toc200886766)

[1.2 国内外研究现状 2](#_Toc200886767)

[12.1遥感图像分割方法 2](#_Toc200886768)

[12.2预训练基础模型与微调方法 3](#_Toc200886769)

[1.3 本文主要工作 5](#_Toc200886770)

[1.4 本文结构安排 6](#_Toc200886771)

[2 相关理论与技术概述 8](#_Toc200886772)

[2.1 遥感图像分割算法 8](#_Toc200886773)

[2.1.1 图像分割定义与场景 8](#_Toc200886774)

[2.1.2 实例分割方法 8](#_Toc200886775)

[2.1.3 语义分割方法 8](#_Toc200886776)

[2.1.4 交互式分割方法 8](#_Toc200886777)

[2.2 深度学习理论基础与网络架构 8](#_Toc200886778)

[2.2.1卷积神经网络CNN架构 8](#_Toc200886779)

[2.2.2 编码器-解码器Transformer架构 8](#_Toc200886780)

[2.2.3循环神经网络RNN架构 8](#_Toc200886781)

[2.3 预训练基础模型理论基础 8](#_Toc200886782)

[2.3.1 基础模型的设计与预训练 8](#_Toc200886783)

[2.3.1 面向语义基础模型 8](#_Toc200886784)

[2.3.2 面向视觉基础模型 8](#_Toc200886785)

[2.4 基础模型微调算法 8](#_Toc200886786)

[2.4.1基础模型的跨领域微调 9](#_Toc200886787)

[2.4.2 低秩微调方法 9](#_Toc200886788)

[2.4.3 适配器微调方法 9](#_Toc200886789)

[2.4.4 提示微调方法 9](#_Toc200886790)

[2.5 分割一切模型SAM理论与应用 9](#_Toc200886791)

[2.6 本章小结 9](#_Toc200886792)

[3 基于SAM自适应微调与特征增强的遥感图像实例分割 10](#_Toc200886793)

[3.1 引言 10](#_Toc200886794)

[3.2 基于SAM的自适应微调与特征增强方法 10](#_Toc200886795)

[3.2.1网络结构与评价模型 10](#_Toc200886796)

[3.2.2 基于自适应适配器与低秩参数实现自适应微调 10](#_Toc200886797)

[3.2.3 基于多通道多尺度的自适应特征增强模块 10](#_Toc200886798)

[3.2.4 多尺度融合编码器设计 10](#_Toc200886799)

[3.3 实验结果与分析 10](#_Toc200886800)

[3.3.1实验参数设置 10](#_Toc200886801)

[3.3.2实验结果分析 11](#_Toc200886802)

[3.3.3消融实验 11](#_Toc200886803)

[3.4 本章小结 11](#_Toc200886804)

[4 基于SAM辅助编码与特征混合解码的遥感图像语义分割 12](#_Toc200886805)

[4.1 引言 12](#_Toc200886806)

[4.2 基于辅助编码与特征混合解码的语义分割方法 12](#_Toc200886807)

[4.2.1 整体结构与评价模型 12](#_Toc200886808)

[4.2.2 基于卷积神经网络的辅助编码 12](#_Toc200886809)

[4.2.3 像素解码器特征嵌入 12](#_Toc200886810)

[4.2.4 多尺度分类预测头 12](#_Toc200886811)

[4.3 实验结果与分析 12](#_Toc200886812)

[4.3.1实验参数设置 12](#_Toc200886813)

[4.3.2实验结果分析 12](#_Toc200886814)

[4.3.3消融实验 12](#_Toc200886815)

[4.4 本章小结 12](#_Toc200886816)

[5 .基于SAM提示微调的遥感图像交互式分割 12](#_Toc200886817)

[5.1 引言 12](#_Toc200886818)

[5.2 基于提示微调与多模态模型的交互式分割 12](#_Toc200886819)

[5.2.1 多模态结构与交互模型 12](#_Toc200886820)

[5.2.2 多模态融合编码 13](#_Toc200886821)

[5.2.3 提示微调方法 13](#_Toc200886822)

[5.2.4 实时交互分割系统 13](#_Toc200886823)

[5.3 实验结果与分析 13](#_Toc200886824)

[5.3.1实验参数设置 13](#_Toc200886825)

[5.3.2实验结果分析 13](#_Toc200886826)

[5.3.3对比实验 13](#_Toc200886827)

[5.4 本章小结 13](#_Toc200886828)

[6 总结与展望 14](#_Toc200886829)

[5.1 本文总结 14](#_Toc200886830)

[5.2 未来工作展望 14](#_Toc200886831)

[致 谢 15](#_Toc200886832)

[参考文献 16](#_Toc200886833)

[附 录 17](#_Toc200886834)

**图表目录**

# 绪论

## 1.1 研究背景及意义

（仅占位，待修改调整）随着最新地理空间数据及其快速提取相关和有用信息的方法的需求显著且日益增长，对精确地理空间信息的需求在全球范围内不断增加，以适应当前的世界需求[1]。在此背景下，基于对象的分析框架应运而生，有助于图像数据的分析。这被称为基于对象的图像分析，旨在将遥感影像划分为有意义的图像对象[2]。图像分析研究的目标显著影响分割算法的选择，通常认为没有完美的算法能够适用于所有情况[3]。图像分割过程中的一项相当苛刻的任务是为不同对象选择分割参数的值，以生成符合用户需求和研究目的的分段。分割质量直接影响后续图像分类过程的性能[4]，预计正确配置的图像分割将导致对景观的更好理解和更高的分类准确性。

而遥感图像分割作为基于计算机视觉的应用中的一个重要组成部分，服务于自主车辆的车道分析[5]和无人机的地理定位[6]。作为图像分割的一种任务，遥感图像实例分割是指在从遥感图像中精确地分离出各个独立的对象，并为每个对象分配一个唯一的标签，同时允许自动化处理大量遥感数据，减少人工标注和分析的工作量，从而可以精确测量地物的面积、边界和形状等属性[7-10]。其高效的自动化处理模式对于土地利用、资源管理等需要高精度测量的应用非常有价值。而语义分割生成对对象的细粒度描绘，嵌入其空间信息，实现像素级别的分割。随着图像分割技术的不断进步，它们已被用于解决多样且数据丰富的遥感问题。遥感图像的语义分割示例包括环境监测[11]、作物覆盖和类型分析[12]、森林中的树种、城市空间中的建筑分类[13]。

早期和传统的遥感图像实例分割方法主要依赖于经典的图像处理和计算机视觉技术。这些方法往往难以应对遥感数据的复杂性和变异性，例如尺度变化、遮挡和光照变化[14]，因此在处理复杂场景和高分辨率图像方面存在一定局限性。而随着深度学习方法的发展，其在许多传统计算机视觉应用、检测和分割中显示出远超传统方法的性能，极大地推动了遥感图像分割的发展。深度学习方法自动提取针对目标分类任务的特征，使其在处理复杂场景时成为更好的选择。然而，深度学习方法通常依赖于大量标记数据，这在遥感图像数据中往往伴随巨量的人工成本[15]，很难为遥感图像中的元素分配有意义的标签。为此，半监督、弱监督和自监督学习等方法应运而生，旨在降低标注成本，提升模型在少量标记数据场景下的表现。特别是大模型的发展显著推动了自监督学习的进步。其展现出的卓越特征提取和表示能力，使其在自然语言处理、计算机视觉等多个领域得以广泛应用。但想要将视觉基础模型扩展到遥感图像领域下，来实现良好的泛化性能，往往面临诸多问题：

**第一，俯视遥感图像与近距离自然场景图像之间存在显著的领域差距。**考虑到遥感图像通常是通过卫星或无人机获取的，具有较大的视角和空间覆盖范围。这些图像往往包含丰富的地理信息，但由于其获取方式和高度，遥感图像的分辨率较低，且受大气条件、光照变化等因素的影响较大。相对而言，近距离自然场景图像则是通过相机在地面上捕捉的，具有更高的分辨率和细节，这使得其强调的是局部特征和细节。因此，二者在图像特征、分辨率、视角和内容上存在显著差异。这种领域差距导致如果仅使用自然图像训练基础模型，可能无法有效地适应遥感图像的语义与位置特征。

**第二，视觉基础模型对复杂背景的遥感图像中多目标、小目标的分割效果不稳定。**由于遥感图像的前景比例远低于自然图像，不同类别之间的前景比例各异，这使得前景-背景不平衡问题更加棘手。在多目标情况下，目标之间的相互遮挡和相似的光谱特征进一步加剧了分割的难度。而小目标在图像中占据的像素较少，容易被背景噪声淹没，且其特征可能更不足以让模型有效识别。因此，视觉基础模型往往在大目标和清晰背景的情况下表现良好，但在复杂背景下，模型的特征提取和分类能力受到限制。

**第三，视觉基础模型针对遥感图像领域不同分割任务的拓展存在局限。**目前视觉基础模型虽然在大量数据下训练得到了强大的特征提取能力，但不同于自然语言处理的基础模型，其往往针对某些特定任务表现良好，而想要将其拓展到具体分割的不同任务，如实例分割、语义分割和全景分割等，往往需要修改网络结构，或者在加入额外的提示与引导的情况下进行重新训练，这给其在遥感图像领域的拓展带来一定的局限性。

因此，针对上述视觉基础模型在遥感图像分割任务中的泛化能力挑战，本课题认为基于视觉基础模型进行微调是一种切实可行的方法。这种微调不仅能够充分利用大模型所具备的强大特征提取能力，还能针对遥感图像的特殊性进行优化。这样的微调策略不仅提升了分割性能，也为当前流行的基础模型在遥感图像领域的应用探索提供了新的视角。随着遥感技术的不断进步，如何将先进的计算机视觉技术应用于实际的遥感数据分析中，将对环境监测、城市规划和资源管理等领域产生深远的影响。因此，深入研究视觉基础模型在遥感图像分割中的微调策略，不仅具有理论意义，也对实际应用具有重要的推动作用。

## 1.2 国内外研究现状

### 12.1遥感图像分割方法

（仅占位 需大改）图像分割是遥感图像分析中的一项关键任务，有助于对图像中每个实例进行语义层面的理解。该过程提供了关于对象的位置、类别和形状的重要信息[16]。对遥感图像中表面的准确感知和理解显著推动了地球观测的遥感技术的进步。这项技术的应用涵盖多个领域，包括但不限于国家安全、土地测绘、灾害监测和交通规划。深度学习算法在遥感图像实例分割中显示出显著的前景，证明了它们能够从原始数据中提取深层次、可辨别的特征[17]

在早期研究中，遥感图像分割往往基于光谱的方法，也称为点-像素图像分割方法，主要集中于对单个像素的分析[18]。其中阈值化是最古老且应用最广泛的图像分割算法之一，它包括图像阈值化和特征空间中的分割。用于分割的阈值算法假设，落在感兴趣图像对象内的像素具有明显不同的属性值。阈值值可以是全局实施或局部实施[19]。考虑到图像中的每个区域在特征空间中构成一个单独的聚类，因而聚类方法可以对光谱特征空间进行划分，单个像素被分配到某个聚类中，而不考虑其邻近像素。随后，需要一个后处理步骤，通过合并具有相同聚类值的邻近像素来创建区域[20]。而基于边缘的图像分割方法通过模仿人类理解图像对象的方式，试图检测区域之间的边缘，并将分段定义为这些边缘边界内的区域[21]。形态学算法[22]在分割遥感影像中被广泛应用。它们将二维多光谱卫星影像视为假设的地形起伏，其中像素的值描绘了该点的高度。随后，相邻像素之间的高度差异可以用来将这些像素分隔为不同的区域，利用图像梯度作为输入，将图像划分为不同区域。随着传统方法的发展，图像分割中的一个趋势是使用基于边缘的算法提取初始段，然后使用基于区域的算法合并相似段。这种混合方法同时考虑了边界信息和相邻区域之间的空间信息[23]。

随着机器学习的发展，深度神经网络迅速吸引了遥感领域的研究兴趣，并转变为广泛使用的深度学习架构。考虑到卷积神经网络最初是为了处理多维数组形式的数据而设计，其在计算机视觉领域特别有前景[24]。这一特定特征使它们成为处理多波段图像的合适工具[25]。深度学习的方法极大的推动了图像分割的发展，因而也衍生出不同的分割目标，包括语义分割、实例分割和全景分割等。

### 12.2预训练基础模型与微调方法

近年来，在大规模广泛数据上训练的基础模型在人工智能领域引发了一场变革性的转变。这些模型一旦训练完成，它们可以作为基础并能够被适应到与原始训练模型相关的广泛下游任务。通过利用广泛的数据集和先进的架构，使其能够捕捉复杂的模式和特征，从而在最少的额外训练下进行特定应用的微调。虽然基础模型的基本组成部分已经存在多年，但最近大型语言模型出现激增的趋势，主要可以归因于数据和模型规模的巨大增加。基础模型被定义为：在自我监督或半监督方式下，基于大规模数据训练的基础模型，可以适应多个其他下游任务。基础模型的范式转变是显著的，因为它允许用更广泛和通用的基础模型替代多个狭窄的任务特定模型，这些基础模型可以一次性训练并迅速适应多种应用。例如，最近具有数十亿参数的模型如GPT-4[44]，已被有效用于零/少量样本学习，在不需要大规模特定任务数据或模型参数更新的情况下取得了令人印象深刻的性能。同样，训练了5400亿参数的Pathways语言模型PaLM[45]在众多具有挑战性的问题上展示了最先进的能力。

与自然语言处理中的大型语言模型并行，最近许多文献中也探讨了用于不同感知任务的视觉基础模型。例如，预训练的视觉-语言模型CLIP[46]在不同下游视觉任务上表现出了有前景的零样本性能，包括图像分类和物体检测。这些基础模型通常使用从网络收集的数百万个图像-文本对进行训练，并提供具有泛化和迁移能力的表征。GLIP模型[47]在人工标注的LVIS和Visual Genome数据集上训练了一个教师GLIP，然后利用它为图像-文本数据预测框，并通过自然语言处理模型检测名词短语。虽然这些大多数模型旨在从公开可用的数据中提取可获取的知识，但最近提出的SAM模型[48]通过构建数据引擎采取了一种创新的方法，在该方法中模型与循环数据集注释共同开发。SAM独特地利用了大量的掩模集合，展示了强大的泛化能力。

而想要将基础模型应用于不同领域，往往需要微调模型来提高在特定任务上的性能，并对模型进行指令调优，使其能够解决不同的下游视觉任务。文献[49]通过仅微调一个线性层，使用特定任务的数据集训练来改善预训练模型在特定任务上的表现。Minderer等人[50]在检测数据集上微调视觉变换器，以创建一个开放词汇物体检测器。通过将视觉数据集转换为指令调优数据集，使得视觉语言模型能够用于下游任务。

遥感中的基础模型的特点在于能够通过自监督学习技术利用大量未标记数据，从而在不需要广泛标记数据集的情况下学习到强大的表征。关键的自监督学习方法包括对比学习，它通过比较同一数据点的不同增强视图来学习表征，以及预测编码，它训练模型从观察到的部分中预测输入数据的缺失部分。SatMAE[51]模型通过预训练和微调过程，利用Masked Autoencoder架构，从而实现对卫星数据的强大表征学习能力。它不仅能够捕捉图像的时间变化特征，还能准确解析不同波段信息，为下游任务提供高质量的模型基础。Scale-MAE方法[52]通过在已知输入尺度下对输入图像进行遮蔽来预训练网络，其中图像覆盖的地球区域决定了ViT位置编码的尺度，而非图像分辨率。其使用标准的ViT主干编码被遮蔽的图像，然后通过带通滤波器解码遮蔽的图像，以重建低频/高频图像在较低/较高尺度上的表示。VIT-AE方法[53]采用约100M参数的简单视觉Transformer，首次提出了旋转可变大小窗口注意力机制，以适应遥感图像的大尺寸和多样化的对象定向问题。通过优化注意力机制，ViTAE能够显著降低计算成本和内存占用，同时提升对象表示的学习效果。这些模型在场景分类、目标检测和变化检测等各种遥感任务中表现出色。尽管取得了成功，基础模型仍面临一些挑战，包括对高质量和多样化训练数据的需求、显著的计算资源以及对特定遥感任务的有效领域适应，解决这些挑战对基础模型在遥感中持续发展的至关重要。

针对遥感图像领域的分割基础模型，通过SAM微调的方式一般被认为是可行且高效的。Ma等人[54]尝试将SAM从自然场景图像适应到遥感图像,通过直接将SAM的输出作为下游任务的独立辅助，仅对SAM结构的小部分进行微调或修改，这限制了SAM解决自然场景图像与遥感图像之间显著领域转移的能力。文献[55]成功地将 SAM 应用于遥感实例分割任务，通过自主生成提示嵌入，解决了SAM依赖提示输入的问题。通过提取编码器的中间层特征以形成提示的输入，并直接生成prompt嵌入，这种设计还避免了从高维到低维再返回到高维特征的梯度流的障碍。文献[56]提出了PointSAM，通过点注释和自我训练框架对SAM进行微调，解决了噪声伪标签和重叠实例等挑战。文献[57]通过在SAM编码器中注入多个并行低秩LoRA结构，并结合金字塔解码器，有效提升了城市人造物体提取的性能。

综上所述，基础模型在遥感图像分割领域的研究正在不断演进，研究者们已经提出了多种有效的策略，涉及图像预处理、特征提取、模型设计及后处理等多个方面。这些方法在应对图像质量差异、标注数据不足以及提升模型泛化能力方面取得了一定的成果。为了进一步推动这一领域的进步，未来的研究可以集中在以下几个方向：首先，持续改进基础模型架构和训练方法，以增强对复杂遥感图像的分割效果；其次，借助无监督、半监督学习和迁移学习等技术，提升在标注数据稀缺情况下的分割能力；接着，开发针对遥感图像特征的自适应分割算法，以提高模型在不同尺度和分辨率下的表现；最后，关注实际应用，推动遥感图像分割技术在土地利用监测、城市规划及环境保护等领域的实际应用。总之，基础模型在遥感图像分割领域仍然面临诸多挑战，同时也蕴含着巨大的发展潜力。通过不断的创新和问题解决，该领域将对遥感技术的未来发展产生深远的影响。

## 1.3 本文主要工作

通过观察国内外的研究现状，可以发现遥感图像分割领域的视觉基础模型微调具有重要的研究价值。因此，本文重点研究基于SAM基础模型在遥感图像中的微调方法。如图所示，本文分别设计了基于SAM自适应微调与特征增强方法与基于SAM辅助编码与特征混合解码方法，并通过提示微调实现基于遥感图像交互式分割。本文的主要贡献总结如下： （待展开）

1. **基于SAM自适应微调与特征增强的遥感图像实例分割**
2. **基于SAM辅助编码与特征混合解码的遥感图像语义分割**
3. **基于SAM提示微调的遥感图像交互式分割**

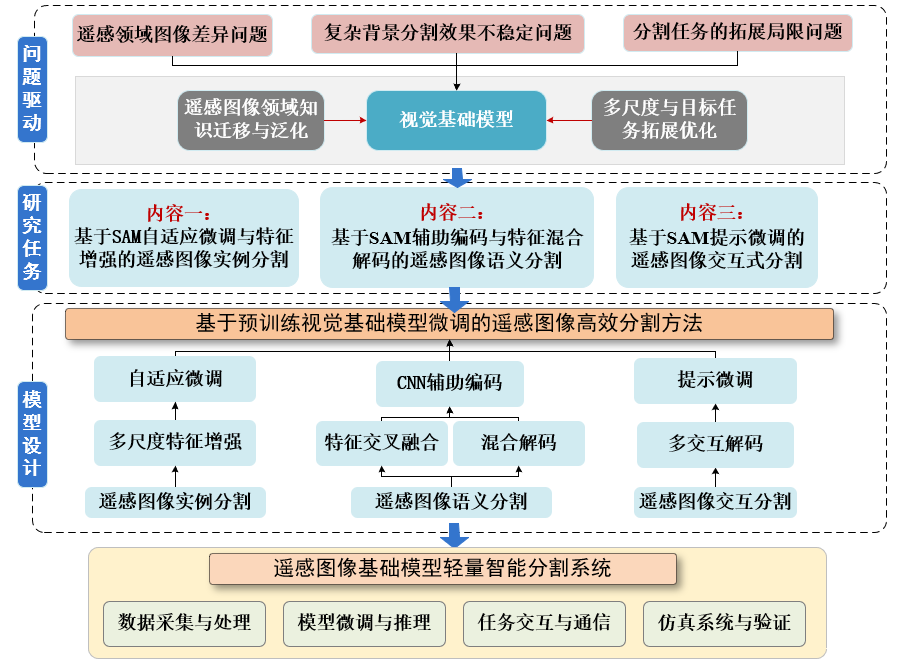


图 本文的主要工作（待简化）

## 1.4 本文结构安排

论文内容一共分为五章，每章的具体研究内容安排如下所示：

第一章为绪论部分，首先介绍了本篇论文的研究背景及意义，介绍了深度学习的发展现状和视觉基础模型的提出背景和目前面对的主要挑战，然后介绍了关于国内外学者对于基础模型微调方法以及遥感图像的分割方法的研究现状，最后介绍了本文的主要工作和内容结构安排。

第二章为相关理论与技术概述部分，主要介绍了图像分割研究工作相关的理论和技术概述，包括深度学习、图像分割算法、预训练基础模型、微调算法以及SAM相关的理论和技术。

第三章介绍了基于SAM自适应微调与特征增强的遥感图像实例分割，旨在解决实例分割场景下俯视遥感图像与近距离自然场景图像之间的领域差距问题。首先通过适配器微调与低秩微调方法结合应用于编码器结构，并通过迭代剪枝的思想实现自适应微调；然后在编码器结构中引入多尺度特征增强模块，用来弥补SAM模型的多尺度特征缺失的弊端；最后通过部分遥感图像对SAM进行微调实验，与多个基准算法进行对比，验证所提方法的有效性。

第四章介绍了基于SAM辅助编码与特征混合解码的遥感图像语义分割，旨在保持SAM原有的泛化能力下，进一步拓展视觉基础模型的应用领域。首先通过辅助CNN编码器，促进SAM对下游任务的适应并补充ViT编码器在密集视觉任务中的效果；然后针对原始SAM的编码器输出，通过像素解码器生成每像素特征嵌入，包含每个像素的全局表示。对于基于每像素分类的分割模型，像素解码器用于直接输出最终结果，以解决遥感图像中目标的稀疏分布导致的前景-背景不平衡问题。最后设计了微调训练实验，将所提算法和多个基准算法进行对比，验证所提方法的有效性。

第五章介绍了基于SAM提示微调的遥感图像交互式分割，通过对提示编码器重新设计，使得可以基于遥感图像实现交互式分割系统；然后对于包含语义信息的提示词，通过基础的预训练语义模型CLIP来生成语义提示嵌入，帮助整体框架更好地理解提取的全局特征；最后将微调的模型作为系统的基础层支持设计一个遥感图像智能分割系统。（待定）

第六章为全文总结部分，对整体的研究内容进行总结并对未来的工作做出展望。

# 相关理论与技术概述

本章主要介图像分割研究工作相关的理论和技术概述，包括深度学习、图像分割算法、预训练基础模型、微调算法以及SAM相关的理论和技术。

## 2.1 遥感图像分割算法

什么是图像分割。

### 2.1.1 图像分割定义与场景

### 2.1.2 实例分割方法

### 2.1.3 语义分割方法

### 2.1.4 交互式分割方法

## 2.2 深度学习理论基础与网络架构

深度学习是xxxx。

### 2.2.1卷积神经网络CNN架构

如图1.1所示，xxx

。

### 2.2.2 编码器-解码器Transformer架构

。xxxxx

### 2.2.3循环神经网络RNN架构

Xxx

## 2.3 预训练基础模型理论基础

### 2.3.1 基础模型的设计与预训练

### 2.3.1 面向语义基础模型

### 2.3.2 面向视觉基础模型

带一句多模态大模型？

## 2.4 基础模型微调算法

### 2.4.1基础模型的跨领域微调

### 2.4.2 低秩微调方法

### 2.4.3 适配器微调方法

### 2.4.4 提示微调方法

## 2.5 分割一切模型SAM理论与应用

什么是SAM 发展有哪些 可以有哪些拓展应用。

## 2.6 本章小结

本章系统阐述了与本研究相关的理论基础和关键技术，构建了一个完整的理论体系，为后续研究工作提供了必要的理论支撑。首先，本章详细介绍了xxx

# 基于SAM自适应微调与特征增强的遥感图像实例分割

## 3.1 引言

近些年来，大模型的发展显著推动了自监督学习的进步。基础大模型如GPT-4[1]、CLIP[2]和DINO[3]在各自领域取得了显著的技术进展，其展现出的卓越特征提取和表示能力，使其在自然语言处理、计算机视觉等多个领域得以广泛应用。基础模型的灵活性和可扩展性使得它们能够在不同领域下，针对特定任务保持相对良好的泛化性能。视觉基础模型Segment Anything Model (SAM) 实现了在包含超过10亿个掩码，覆盖了1100万张的图像数据集上的模型训练，可以为任何图像或视频中的任何物体生成实例分割掩码，同时无需额外的训练，具有十分优秀的泛化和迁移能力。

为了进一步提升基础模型SAM的在特定应用场景下的性能提升，参考语言序列基础模型的微调技术，基于SAM的模型微调也被广泛研究并应用。HQ-SAM[4]设计了一个可学习的高质量输出令牌，通过注入到SAM的掩码解码器中来预测高质量的掩码。FastSAM[5]和MobileSAM[6]都是对SAM的轻量型改编，旨在提高分割效率和在资源受限设备上的性能，并支持其他领域的训练与应用。SAM\_MLoRAF[7]通过在SAM编码器中注入多个并行低秩LoRA结构，并结合金字塔解码器，有效提升了城市人造物体提取的性能。但由于遥感图像数据的特殊性，想要将分割一切大模型 (SAM) 直接应用于遥感图像的实例分割任务仍有许多问题。首先，SAM严重依赖于人工先验提示，需要手动输入的点、框、掩码和文本提示，且分割结果没有标记类别；其次，SAM对目标对象的细节部分，特别是对于遥感图像的多目标、小目标的分割容易出现错误；最后SAM对于遥感图像数据的特征提取和实例分割能力泛化型不足，缺乏相应的先验知识。

为了解决这一挑战，本文通过在分割一切大模型的骨干框架中引入可迭代调整参数的适配器微调模块，使得微调模型可以适应遥感图像下游任务。适配器模块中通过评分函数计算不同神经层的重要性，实现动态且多尺度的适配器参数分配。然后，一方面从多个适配器模块的中间层进行特征提取与聚合，得到微调后的全局多尺度语义信息特征；另一方面，从原始基础模型的掩码解码器中提取初始掩码特征，最终通过卷积处理对二者进行总结，得到高质量掩码特征。此外，本文还利用基于锚点的区域建议网络生成候选目标框，通过语义分支和定位分支得到目标的位置提示，输入到掩码解码器中实现全自动的先验提示，以解决人工提示的问题。

本部分的主要贡献包括以下几个方面：

1）设计了一种自适应的微调范式，通过结合自适应低秩微调AdaLoRA[8]和插入适配器模块MiMi-Adapter[9]的思想来优化SAM在遥感图像上的泛化性能，并修改了编码器的输出特征，使其聚合了微调后的全局特征以及多尺度特征；

2）引入了一种自适应特征融合增强的方法，用于融合不同尺度的特征信息，以优化分割掩码，使其具有凸显丰富纹理或物体边缘的性能；

3. 我们针对本文提出的自适应微调与特征融合方法在两个公开遥感图像数据集中进行了广泛验证，实验表明了我们的方法在遥感图像实例分割的有效性与泛化性。

本章的组织结构如下：在3.2节中，我们将分别介绍针对遥感图像实例分割的微调SAM网络框架、自适应微调范策略、全局特征聚合以及自适应多尺度特征增强方法。在3.3节中，本章对所提出算法在两个公开数据集上进行了对比试验以及消融实验，并对实验结果进行了分析和讨论。最后，3.4节是对本章内容进行的总结。

## 3.2 基于SAM的自适应微调与特征增强方法

### 3.2.1网络结构与评价模型

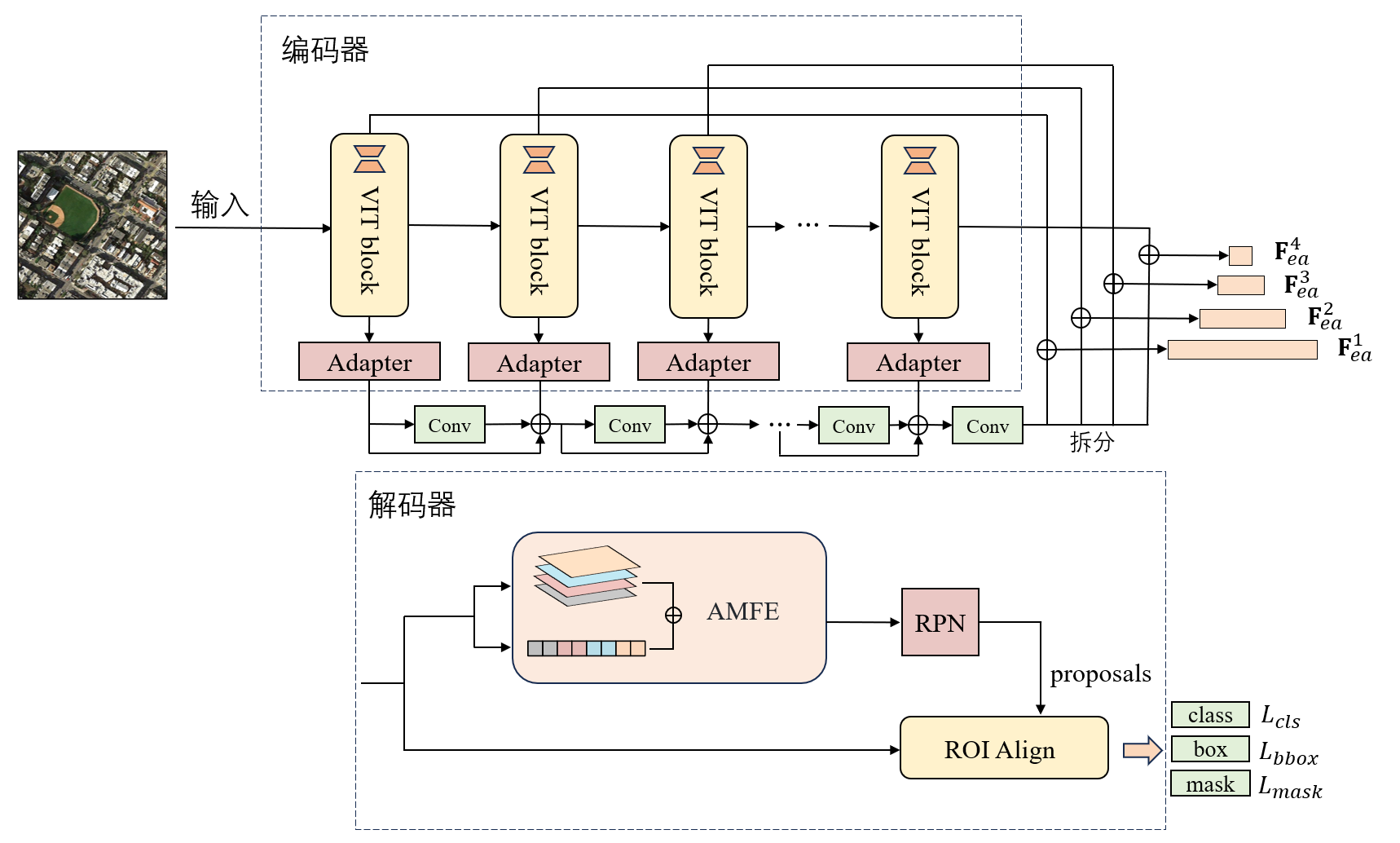


图3.1 基于SAM的自适应微调的遥感图像分割网络结构图

本章所提出的基于SAM的自适应微调与特征增强方法网络结构如图 3.1所示。该网络仍然采用SAM的编码器-解码器结构，主要包含三部分内容：针对SAM编码器的自适应微调模块、全局多尺度聚合器和解码器中的自适应多尺度增强方法。在编码器中，每个Transformer块都应用了AdaLoRA的微调方法，并有选择性的插入了MiMi-Adapter以进一步学习参数来微调SAM模型，来增强SAM在遥感图像上的泛化能力。本方法还修改了SAM编码器的输出，通过一种全局多尺度特征相加的方式得到金字塔结构的特征，将其传入到修改后的解码器中，并舍弃了SAM解码器中的提示编码器，重新设计了mask解码结构。将金字塔特征进行增强之后，输入到区域提议网络，生成候选区域，并通过ROI Align模块进行对齐，最终用于分类、边界框回归和掩码计算，以实现自动实例分割。

我们采用了标准的 COCO 平均平均精度（mAP）来定量评估不同方法的实例分割结果。通过计算不同IoU阈值下的平均精度来评估模型性能:

其中，IoU是指预测掩模与真实掩模的交集面积与它们的并集面积之间的比率，是指从0.5到0.95，步长为0.05的一组IoU阈值，mAP是指在10个掩模交并比阈值下的平均精度，AP是Precision-Recall曲线下的面积，越大的AP值表示预测的实例分割掩码越准确。

### 3.2.2 基于自适应适配器与低秩参数实现自适应微调

考虑到SAM模型本身具有强大的特征提取能力，我们保持SAM原有的结构，同时冻结视觉编码器，引入自适应低秩微调AdaLoRA[14]并选择部分transformer块插入了自适应适配器MiMi-Adapter[15]：

**（1）自适应适配器SAM-MiMi-Adapter：**

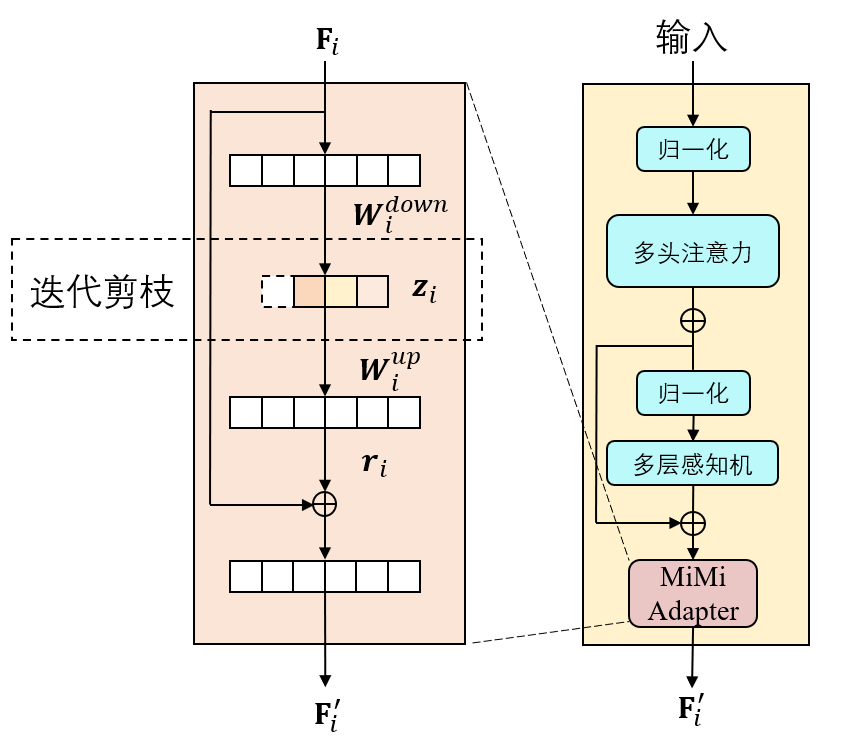


图3.2 MiMi适配器在Transformer层中的结构

为了增强SAM在遥感图像的迁移能力，我们还将自适应微调模块MiMi-Adapter[15]插入到SAM的标准Transformer块中，如图3.2所示，为了降低参数量，我们舍弃了原方法中在多头自注意层引入的Adapter模块，并将插入位置修改为每个变压器块中的MLP层残差连接之后。第个适配器的输入为。适配器首先通过一个全连接层将 降维到，并使用激活函数 。这个层由矩阵 参数化。然后，第二个全连接层将 转换回输出 ，由矩阵 参数化。最后，适配器使用残差连接，使得输出 的计算如下：

其中，和均为线性投影矩阵，分别用于上采样和下采样，为输入向量，为输出向量。

MiMi-Adapter采用了迭代剪枝策略，即使用随机梯度下降的方法，通过最小化交叉熵损失来学习适配器参数 。这是包含了每个第i个 ViT 子层的适配器参数的集合。然后，根据一个评分函数衡量每个适配器神经元的重要性，该评分函数根据中相应行的范数和中相应行的范数之和计算得出：

**（2）自适应低秩微调方法SAM-AdaLoRA**

LoRA[7]是目前广泛用于训练大型语言模型的一种微调，通过在原特征一侧添加一个低秩的矩阵来实现模型在下游任务上的微调。但由于 LoRA没有考虑到模型中不同深度特征的重要性，因此AdaLoRA[2]提出了基于SVD的形式参数化增量更新，这种方法可以在规避SVD复杂的计算的同时高效裁剪不重要的奇异值，从而降低计算量。本章提出的方法在图像编码器中的每个变换器QKV（Query, Key, Value）层引入AdaLoRA，并在训练步骤中采用SVD的形式将增量矩阵参数化更新。

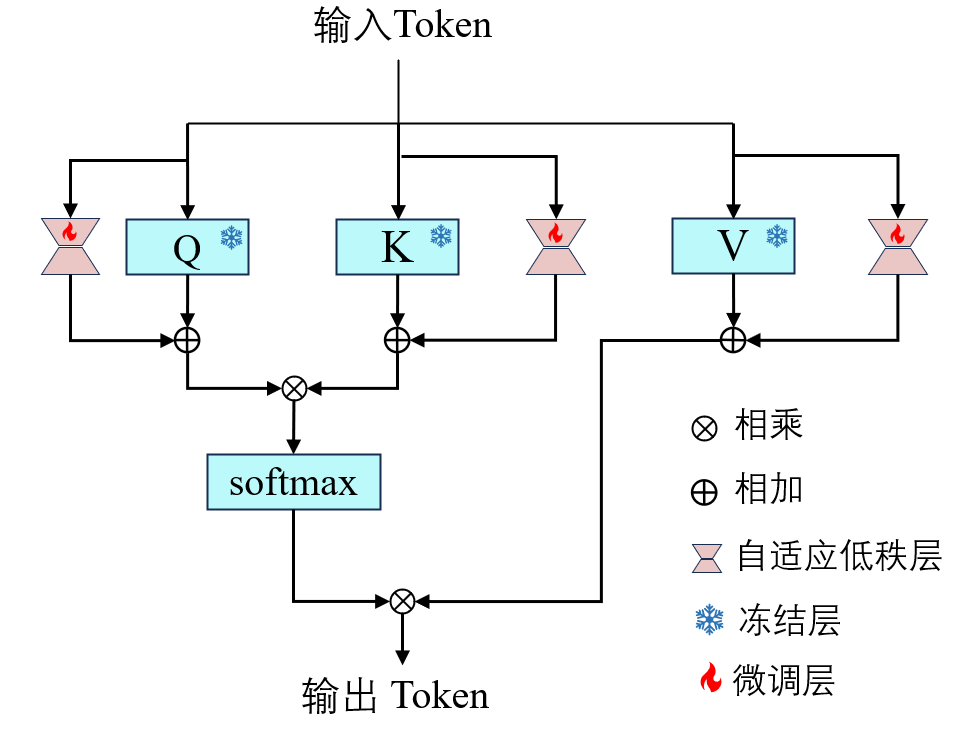


图3.3 Transformer层中的自适应低秩模块

如图3.3所示，我们在图像编码器中的每个变换器QKV（Query, Key, Value）层，训练步骤中采用SVD的形式将增量矩阵参数化更新：

其中，是初始权重，是权重的增量更新，被参数化为SVD的形式：，其中和都是低秩矩阵，维度分别为d×r和r×d，是一个对角矩阵，表示不同特征的权重，维度为r×r。d是输入特征的维度，r是选定的低秩。

然后，基于参数分配的重要性，对冗余的奇异值进行剪枝，以调整低秩矩阵的秩，重要性分数由下式给出：

其中是任意可训练的权重, 表示前t个时刻该单参数重要性的平滑值，计算敏感性的局部时间变化并进行滑动平均。

### 3.2.3 基于多通道多尺度的自适应特征增强模块

在解码器中，本文设计了一种称为自适应多尺度特征增强的方法AMFE，使其能够自适应地增强不同尺度的特征。通过在空间维度和通道维度上应用加权调整，可以增强层间和层内的特征。

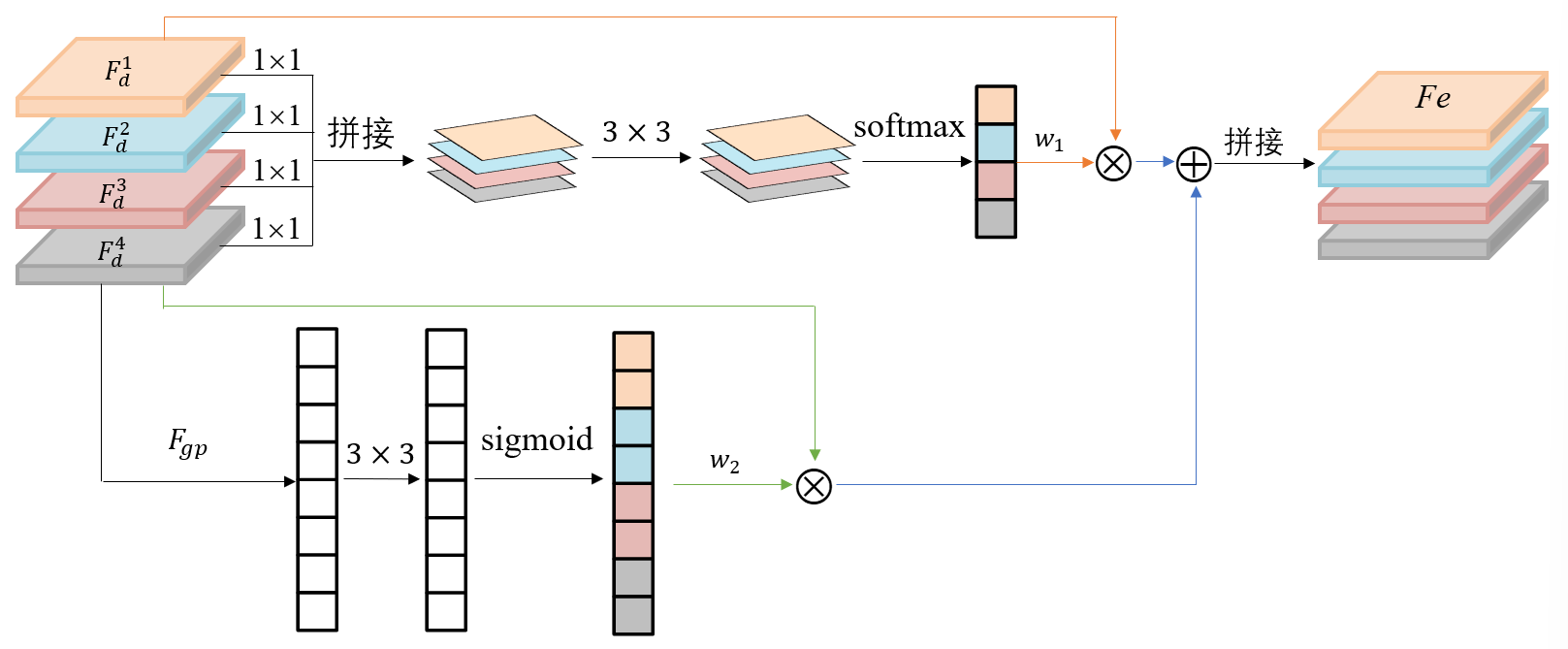


图3.4 自适应多尺度增强模块结构图

如图3.4所示，经过统一维度尺寸处理的输入的特征记为，其中，和是空间尺寸，是通道数。

在通道维度上：首先对输入特征应用1×1的卷积操作，得到一组特征图。将这四个特征图进行拼接，得到新的特征表示，在对齐应用3×3的卷积，以增强空间信息。然后对上述卷积结果采用softmax函数得到空间维度的自适应权重经过一个1×1卷积并拼接之后，使用一个3×3的卷积来增强空间信息，随后通过softmax处理得到其自适应权重; 将权重乘以原始特征，实现空间维度的加权增强：

其中表示逐元素乘法，表示空间维度增强的特征。

在空间维度上：首先将Fdi通过全局平均池化将其压缩为单尺寸，得到其通道描述：

其中为池化后的特征，GAP表示全局平均池化操作。

随后同样使用一个3×3的卷积并经过sigmoid处理得到通道维度的权重，将其乘以原始特征，实现通道维度的加权增强：

最终，将空间维度增强特征和通道维度增强特征相加，得到自适应增强后的特征:

其中。与MaskRCNN[8]方法保持一致，我们将处理之后的增强特征图作为RPN的输入进行后续的预测。

### 3.2.4 多尺度融合编码器设计

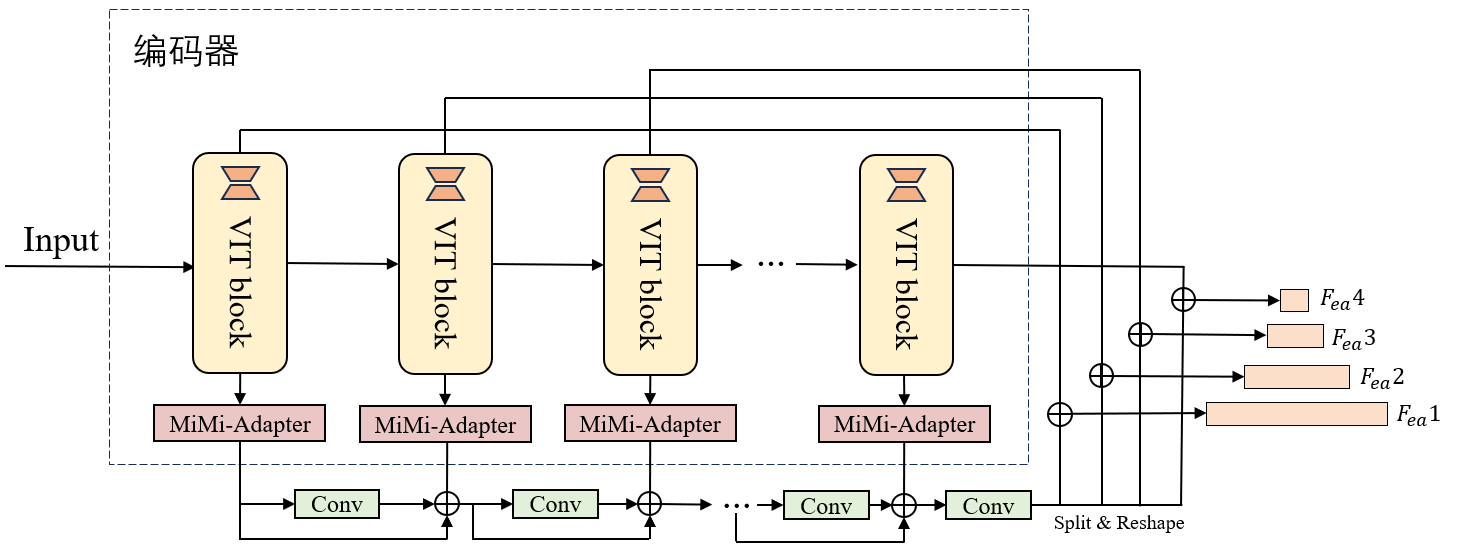


图3.5多尺度特征聚合

为了增强编码器输出特征的多尺度信息并聚合微调后的特征，本文采用了如图3.5所示的特征组合方法。整体过程可以用以下方程表示：

其中，是MiMi-Adapter的输出特征，该模块通过一个3×3卷积-ReLU块增强了空间信息。随后，利用残差连接将这些特征聚合得到。接着，采用转置卷积对进行特征分割，以获得不同尺度的特征：

这些特征被依次加到原始 Transformer 模块的输出特征上，得到由AdaLoRA微调得到的特征，从而最终得到多尺度金字塔结构特征，并将其输入解码器进行后续处理。

### 3.2.5分割掩码生成与损失函数

在解码器部分，解码器接收来自编码器的特征表示，以及经过分类和边界框回归模块处理后的候选框信息。通过转置卷积实现上采样，逐步恢复空间分辨率。对每个候选框区域，解码器输出一个二维掩码概率图，表示该区域内每个像素属于该实例的概率。生成的掩码概率图会与真实掩码标签进行对比。采用二元交叉熵损失逐像素计算预测掩码和真实掩码之间的差异，从而指导模型学习更准确的掩码边界。

由于解码器结构与 Mask R-CNN [11] 保持一致，所提出的方法主要包含三部分损失：分类损失、边界框回归损失和掩码损失。具体来说，交叉熵损失用于计算目标类别的差异，平滑 L1 损失用于评估预测边界框与真实边界框之间的差异，二元交叉熵损失用于评估实例分割的掩码，其中边界框的 IoU 用于确定监督匹配准则。损失函数的计算可描述如下：

其中，是样本数量，是类别数。是真实标签，表示样本属于类别。表示平滑L1损失。是模型预测的概率，是真实边界框的坐标，是第个实例在第个像素位置上的预测掩码值，是预测边界框的坐标。和分别是掩码的高度和宽度，是用于平衡各损失项权重的超参数。

## 3.3 实验结果与分析

本节将实现提出的基于SAM自适应微调与特征增强的遥感图像实例分割算法，并且进行了广泛的实验以证明算法的有效性。下面将从实验设置、结果分析与消融实验等方面进行介绍。

### 3.3.1实验参数设置

本章在两个公开的数据集上验证所提方法的有效性： NWPU VHR-10数据集[]和SSDD数据集[]，前者主要用于多类别分割任务，后者主要针对小目标检测和分割，他们在大小、类别和场景上各不相同，在遥感实例分割方面得到了相对广泛的应用。以下是数据集的详细介绍：

（1）NWPU VHR-10 数据集是一个用于高分辨率遥感图像分析的公开数据集，主要用于目标检测和图像分割等任务。它包含来自不同遥感卫星的高分辨率图像，涵盖了多种地理区域和场景，包含 10 个主要类别：飞机、船舶、储罐、棒球场、网球场、篮球场、田径场、港口、桥梁、车辆。图像大小在长度和宽度上范围从 500 到 1100 像素。该数据集将图像分为正图像集和负图像集。正集包含一张或多张含有地面目标的图像，而负图像集包含没有地面目标的图像，分别总计 650 张和 150 张。实验将80%的数据用于训练，剩余的20%用于测试。

（2）SSDD船舶检测数据集是一个专门用于合成孔径雷达（SAR）图像中船舶检测任务的公开数据集，SAR 图像一共1160张，分辨率范围从1到15米不等，包括了2540个船舶实例，并涵盖了多种海洋环境和天气条件，提供了丰富的场景变化。实验将80%的图像分配用于训练，其余20%用于测试。

本章将所提方法与当前一些先进的实例分割方法进行了比较，包括传统二阶段方法MaskRCNN，还包括基于SAM的方法Rsprompter和MC-SAM。同时，本章还将SAM原始模型加入对比实验，通过预训练的FasterRCNN模型来生成提示框作为SAM的输入以实现自动提示，Rsprompter则方法采用原论文中的基于Anchor的方法，MC-SAM[30]则使用Mona适配器来微调SAM，以此来对比不同适配器对微调的影响。

本章实验采用了以ViT-B为骨干的SAM预训练模型，图像编码器在训练过程中保持冻结，图像输入大小和SAM保持一致，都为1024×1024。为了提高效率，实验中将每三个Transformer块后的特征提取用于聚合，并使用AdamW作为优化器，权重衰减值为0.01，初始学习率设置为1e−4，并在训练过程中使用余弦退火策略进行衰减。两个数据集的输入批次均设定为8，对不同方法均指定训练100个epoch来进行对比。所有实验均在单个NVIDIA GTX 3090 GPU上使用PyTorch框架实现。本实验中使用3.2.1小节中所提的、和进行评估，并对上述模型重新实现，确保尽可能接近他们的原始论文。

### 3.3.2实验结果分析

**（1）定量结果**：

表3.1和表3.2展示了不同模型在NWPU数据集与SSDD数据集上的定量结果。实验证明，基于SAM自适应微调与特征增强的遥感图像实例分割方法在各项指标上均表现出色。具体而言，本章方法在NWPU的数据集上达到了68.9的mAP分数，均优于其他方法；在SSDD的数据集上达到了71.2的mAP分数，仅次于rsprompter方法，这可能由于即使加入了部分微调方法，SAM本身对合成孔径雷达数据集这样背景噪声多、边界不明显的图像特征仍不敏感。值得关注的是，本方法在mAP-50和mAP-75两个更严格的评估指标上均领先于其他方案，且相较于同样基于SAM微调的MC-SAM方法（其引入Mona模块进行微调），本方法表现出更优的准确率和分割效果，充分说明所设计的微调策略在遥感领域具有更强的适应性和泛化能力。与此同时，未经微调的原始SAM-base模型由于缺乏针对遥感图像的领域知识迁移，表现不佳，难以满足实际应用需求。此外，相较传统的Mask R-CNN方法，微调后的SAM模型凭借其强大的特征提取能力显著提升了解码器的实例分割效果和精度，进一步验证了微调策略和特征增强对提升遥感图像实例分割性能的重要作用。

表3.1 不同模型在NWPU数据集上不同IoU阈值下掩码的平均精度

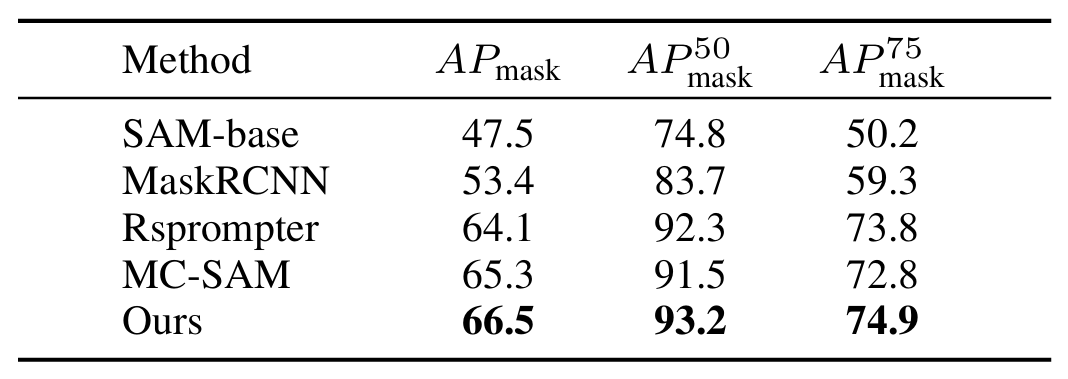


表3.2 不同模型在SSDD数据集上不同IoU阈值下掩码的平均精度



**（2）定性结果：**

为了更直观且全面地比较各方法的实验表现，图3.6、图3.7图、3.8、图3.9分别展示了不同实例分割方法在NWPU数据集和SSDD数据集上的定性结果。从图中可以明显观察到，未经微调的原始SAM模型即使在输入较为准确的提示条件下，其在遥感图像实例分割任务中生成的掩码仍显得不够清晰且完整，尤其是在目标边界模糊、背景噪声复杂的区域，原始SAM的细粒度分割能力表现较弱，难以准确区分目标与背景。相比之下，采用双阶段分割策略的Mask R-CNN方法在两个数据集上的检测和分割效果均不理想，这一现象可能源自其训练迭代次数不足以及数据集规模有限，导致模型难以充分学习复杂的遥感图像特征。进一步对比基于SAM的各类改进方法，本文发现经过自适应微调的模型显著提升了分割结果的边缘锐利度和轮廓清晰度，相较于原始SAM-base模型，微调后的模型更能准确捕捉目标形态细节。然而，基于提示微调的Rsprompter方法在处理目标重叠区域时表现欠佳，难以有效分离相邻目标，而MC-SAM方法则存在对目标位置敏感度不足的问题，导致定位精度有所下降。综合来看，本章提出的基于SAM自适应微调与特征增强的实例分割方法不仅在目标定位上表现出更高的准确性，而且生成的掩码质量更接近真实标注。

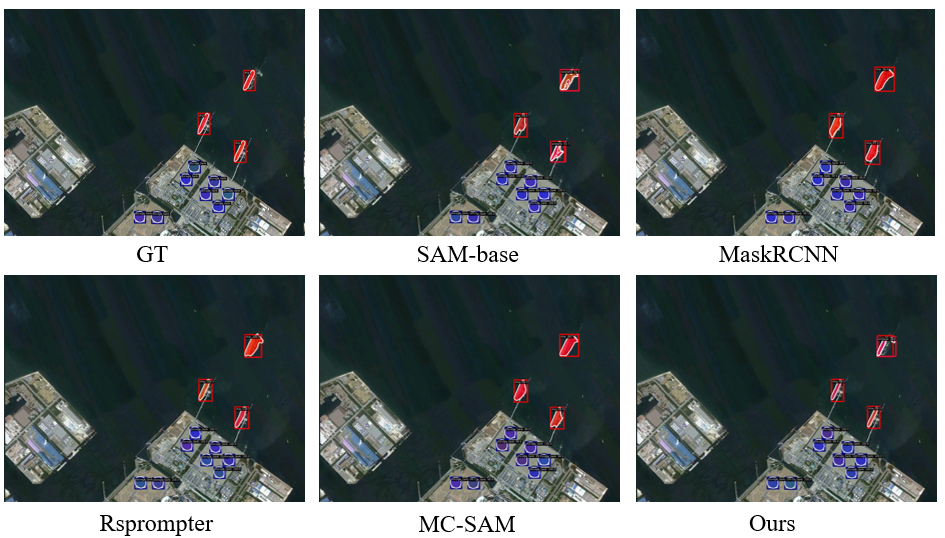


图3.6 不同方法在NWPU数据集上的可视化结果(1)

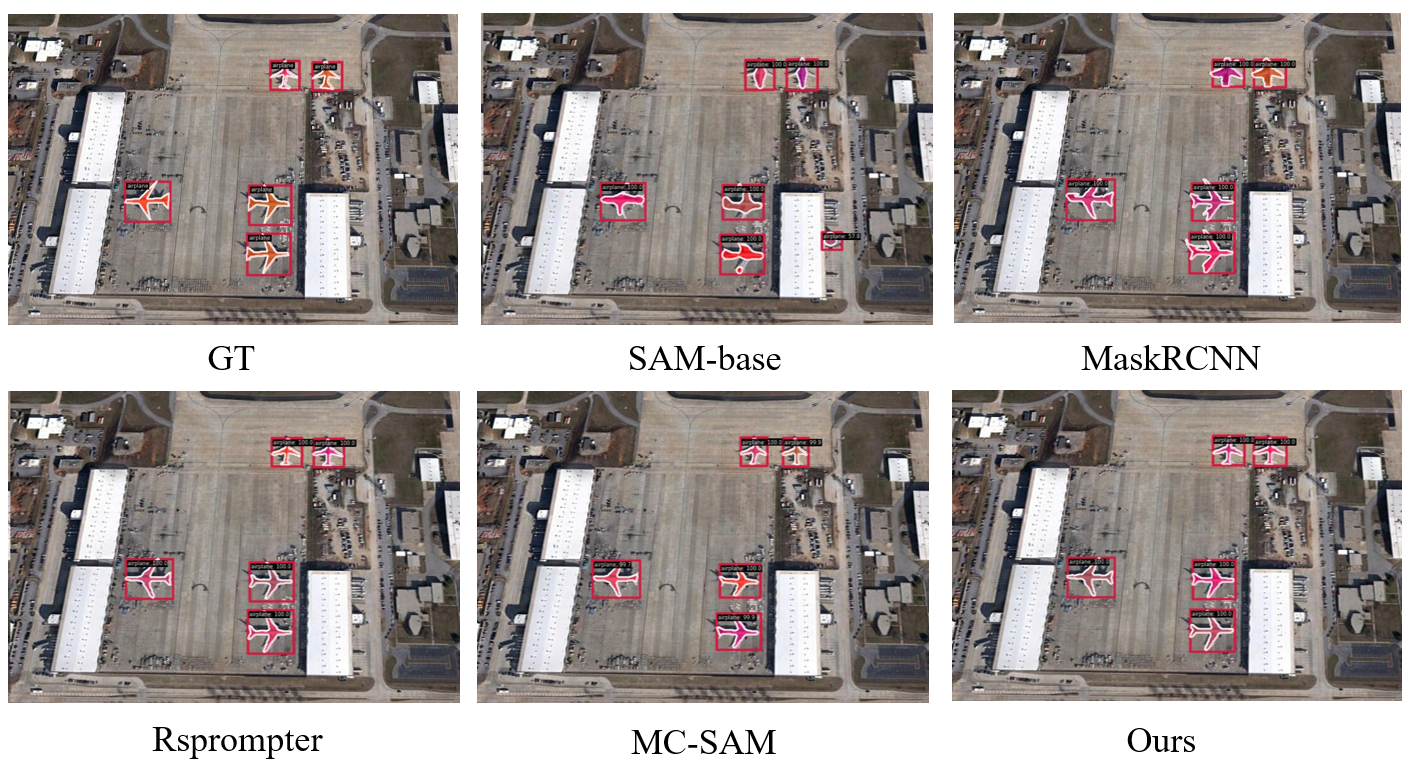


图3.7 不同方法在NWPU数据集上的可视化结果(2)



图3.8 不同方法在SSDD数据集上的可视化结果(1)

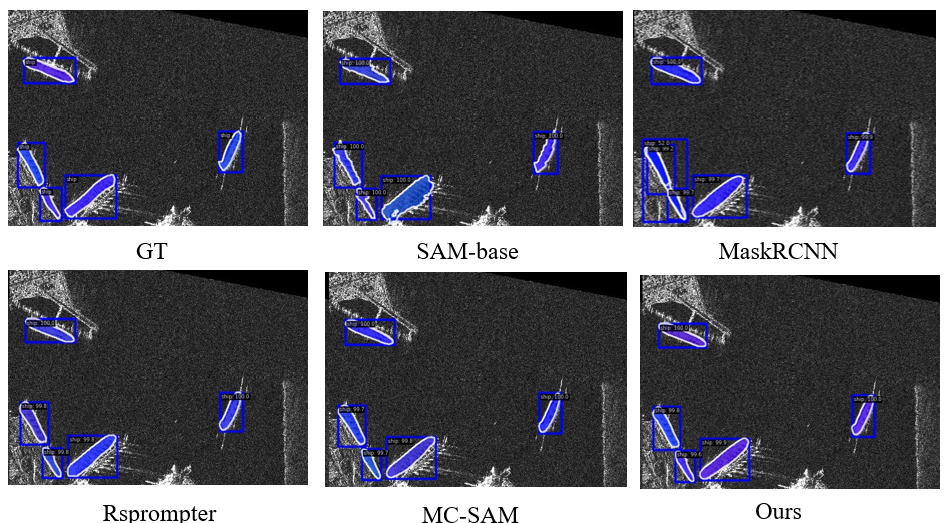
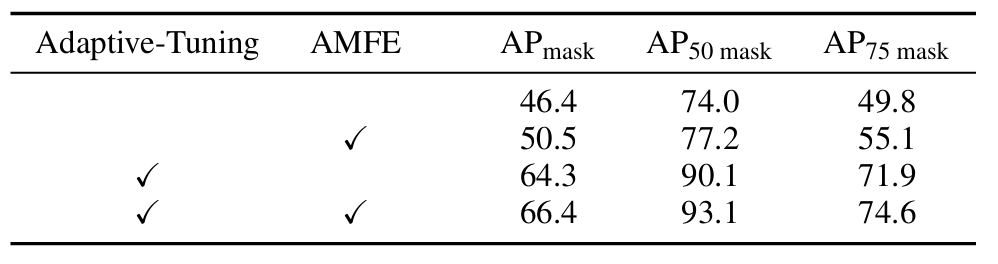


图3.8 不同方法在SSDD数据集上的可视化结果(2)

### 3.3.3消融实验

为了系统性地验证本文所提出的自适应微调方法以及多尺度特征增强模块（AMFE）的有效性，本文在NWPU数据集上设计并开展了一系列消融实验，针对所提模型的关键结构组件进行了深入分析。表3.3详细展示了不同配置组合下的实验对比结果，量化了各模块对模型性能提升的贡献。具体来看，基线模型即原始SAM方法在未做任何改动时的表现作为参照。引入多尺度特征增强模块AMFE后，模型在不同IOU阈值（如0.5、0.75及平均IOU）下的掩码平均精度（AP）均实现了显著提升，分别提高了4.1%、3.2%和5.3%。这表明AMFE模块有效增强了模型对遥感图像中多尺度目标的特征表达能力，提升了分割的细节捕捉能力。进一步地，近通过对SAM模型的编码器进行自适应微调，使模型能够更好地适应遥感图像的特性，掩码的AP值在各个IOU阈值上实现了更加显著的跃升，分别提升了17.9%、16.1%和22.1%，表明微调策略对模型性能起到了很好的提升。当将自适应微调与AMFE模块相结合时，模型在掩码分割精度上达到了最优表现，AP值分别提升至20.0%、19.1%和24.8%。综上所述，实验结果明确验证了本文提出的各个模块在提升遥感图像实例分割性能方面的积极作用，体现了该方法的有效性。

表3.3 不同微调模块作用下的分割准确率



## 3.4 本章小结

由于到遥感图像数据的特殊性，直接将分割一切大模型 (SAM)应用于高光谱图像的实例分割会导致分割效果不稳定，因为SAM不仅严重依赖于人工先验提示，而且对目标对象的分割细粒度不足。在考虑到SAM对于遥感图像数据的特征提取和实例分割能力泛化型与先验度的情况下，本章方法结合了自适应低秩和可变适配器模块提出了一种SAM的微调方法，增强了编码器输出特征的全局尺度信息。通过舍弃SAM原有的提示编码器和解码器，设计了一个自适应多尺度特征融合的解码器来实现自动实例分割。在此基础之上，本章利用两种公开遥感图像数据集对SAM模型进行了微调，并与现有方法进行了对比，不仅证明了此方法对SAM模型微调的有效性，还表现出更高质量的掩码分割效果。该方法未来也可以推广到更多领域的基础视觉任务。

# 基于SAM的CNN辅助编码与特征混合解码的遥感图像语义分割

## 4.1 引言

虽然SAM作为视觉基础模型在大多数场景下表现出卓越的泛化能力和零样本能力，但下游任务的复杂需求和多样的应用领域对SAM提出了更复杂的要求和进一步的优化。尤其对于合成孔径雷达（SAR）图像[2]，[3]，SAR的电磁波成像机制独特，导致其图像特征具有特殊性，而这对于SAM来说是陌生且难以解读的。近年来，一些基于深度学习的工作旨在提升传统模型在SAR图像分割任务中的表现。FFSSNet[4]旨在实现不同尺度SAR图像的特征融合；LGMFNet[5]为SAR建筑物分割任务提供了双编码器-解码器架构；MAZ-Net[6]则探索了双极化SAR图像上的海洋水产养殖区提取算法。然而，SAR图像的多样分辨率和不同极化方式往往使SAM在像素级分割时产生混淆，因为SA-1B数据集的预训练数据无法覆盖SAR图像分割任务中的此类细粒度情况。SAR图像由于电磁波成像机制引入的干扰特性以及某些低信噪比（SNR）场景，与自然场景存在显著差异且过于独特，难以在训练自然视觉基础模型时加以考虑。这导致SAM在SAR图像分割任务上的效果和可靠性因此受损[7]。另一方面，由于获取SAR图像的机会稀缺且成本高昂，基于有限的SAR图像数据资源从零构建SAR基础模型较为困难。利用SAM进行SAR图像解译呈现出值得探索的可能性，但同时也面临以下两大主要挑战。

SAM几乎无法对SAR图像进行合理精度的分割，尽管其在包括光学航空影像在内的自然图像上表现出良好的泛化能力，如图1(a)所示。考虑到SAR信号回波的电磁波成像机制，SAR图像的表现与自然场景图像有很大不同。具体而言，由于波动物理、旁瓣及复杂的散射特性，SAR图像表现出相干斑点噪声，这可能引入误警报，以及在陡峭地形或高大建筑区域出现重叠效应和阴影效应。SAR图像的特定情况，包括多极化特性和低信噪比成像环境，在SAM中很少涉及。2）SAM的类别无关特性阻碍了其在多样化视觉下游任务中的应用。对于SAR图像的语义分割和土地覆盖分类任务，每个像素都需要预测类别，而SAM仅在提示输入的指导下生成若干二值掩码，缺乏类别信息。正如图1(b)所示，SAM对SAR图像的分割结果几乎不包含有意义的语义信息，因为同一类别的像素未被归类为同一类别。此外，土地覆盖分类任务要求模型对每幅图像输出多类别预测掩码，而SAM输出的掩码数量可变。因此，我们提出了一种名为classwise-SAMadapter（CWSAM）的方法，通过利用基础模型SAM并添加类别轻量级掩码解码器，实现了像素级SAR土地覆盖分类任务的合理性能。采用了精心设计的adapter微调机制，在冻结SAM中重量级视觉变换器（ViT）[8]图像编码器的大部分参数的同时，通过adapter插入少量可训练参数，CWSAM将SAM的自然场景领域迁移至SAR领域，探索大型模型在SAR下游任务（如土地覆盖分类）中的潜在应用。由于模型权重中仅有限参数可通过梯度训练，CWSAM需求较低计算资源，训练过程高效。我们构建CWSAM作为视觉基础模型应用于SAR图像的初步尝试，并相信此类具备可行计算成本的有效方法，将为基础模型在SAR智能解译的发展带来良好前景。大量实验表明该方法在分割性能和训练过程上的效率与可靠性。CWSAM以更少的可训练参数和更低的内存消耗，优于多种最先进语义分割算法。据我们所知，这是首次将SAM引入SAR领域进行参数高效微调，以实现具备类别感知能力的语义分割下游任务。综上，本文贡献可总结如下：1）提出的CWSAM引入端到端架构，在SAM图像编码器中插入轻量级adapter，实现大模型的高效参数微调。该方法将SAM的自然场景领域迁移至SAR领域。通过利用视觉基础模型，在SAR图像的土地覆盖分类任务中实现可靠性能。2）设计了CWSAM的类别掩码解码器，将类别属性扩展至原类别无关的SAM模型，支持精细语义分割下游任务。针对SAR图像，类别掩码解码器中轻量级的类别级上采样卷积层用于像素级识别多类别土地覆盖。3）设计了SAR图像低频信息的任务特定输入模块，通过二维图像快速傅里叶变换在端到端架构中自动提供充分的土地覆盖语义信息，提升分割性能。全面实验表明CWSAM超越多种现有语义分割算法，同时无需高计算成本。尽管面临SAR与SAM自然图像特性差异，所提方法成功利用基础模型优势提升SAR图像的分割性能。

然而，基于该范式的方法通常因自然图像与遥感图像间的差异而表现有限。因此，最近提出了一些大规模遥感模型[9]，[10]以解决此问题。尽管如此，这些遥感大规模模型主要强调基础模型的通用预训练，而非针对具体分割任务，亦限制了分割性能。值得注意的是，基于前述在自然图像上进行通用预训练的模型也存在此限制。

通用用途也存在这一限制。因此，为了避免上述限制，我们引入了一种基于自然图像预训练的分割任务范式，然后对遥感（RS）图像进行微调。尽管这对于小规模模型并非新范式，但目前仍缺乏基于该范式的遥感语义分割大规模模型方法。这是因为两个挑战限制了大规模自然图像分割模型，尤其是“分割一切模型”（SAM）[11]，在遥感图像语义分割中的适应性。首先，如王等人[12]所证实，许多大规模视觉模型（即视觉变换器）的主干网络具有低通滤波特性，这可能导致分割边缘不清晰。第二，天然图像和遥感图像的特性不同。实际上，遥感图像中存在许多不同尺度且变化较大的物体，而自然图像中物体的尺度变化通常要小得多。因此，大规模自然图像分割模型难以适应遥感图像中多样的物体尺度和不均匀分布。尤其对于SAM，除上述限制外，其输出tokens是针对自然图像训练的，鉴于自然图像与遥感图像的分布差异，直接用于遥感图像的语义分割并不合适。RS-Prompter [13]和Osco等人[14]分别通过添加额外的提示生成器和输入遥感图像提示解决了这一问题。因此，提示数据集对于他们的方法来说是不可或缺的，而制作提示数据集耗时且资源消耗巨大。此外，[13]和[14]中提出的机制是针对遥感实例分割而非语义分割。为缓解上述问题，我们提出了一种多尺度增强SAM（MeSAM）用于遥感图像的语义分割。具体而言，我们的方法以SAM [11]作为遥感语义分割的基础模型。得益于拥有大量带掩码的训练数据，SAM通过提示学习进行分割任务预训练，而非用于通用目的的自监督学习[15]，这赋予了该模型卓越的分割泛化能力。另一方面，虽然SAM出色的泛化能力使其能够直接分割遥感图像，但因上述问题，其性能仍受限，如图1所示。因此，我们首先引入了inception混合器重构SAM图像编码器中的注意力模块，形成inception变换器，从而避免了视觉变换器低通滤波导致的高频信息丢失。随后，我们通过多尺度卷积核和来自inception变换器的跳跃连接实现多尺度特征提取，从而使原始SAM适应具有变尺度物体的遥感图像。最后，通过在解码器中加入可学习的遥感token及校正模块来修正自然图像的输出tokens，缓解了自然图像与遥感图像分布差异带来的性能限制。本文贡献总结如下：1）我们提出了一种用于遥感图像语义分割的大规模模型MeSAM，其性能优于已有的遥感语义分割最新先进方法（SOTA）。2）针对SAM，提出了一种高频和多尺度特征增强的高效微调范式，填补了基于分割预训练然后在遥感图像上微调的大规模语义分割的空白。与低秩适配（LoRA）[16]和适配器[17]等SOTA高效微调方法相比，所提微调方法取得了更佳结果（详见第四章D1节）。3）引入inception混合器微调SAM中的图像编码器，提高了高频和多尺度语义特征提取能力，提升了不同尺度以及具有丰富纹理或边缘物体的分割表现。4）在SAM的掩码解码器中设计可学习token和校正模块，用于遥感语义分割，可无需任何提示微调即校准自然图像输出tokens至遥感图像。由此，我们可以省去遥感语义分割提示数据集的收集。本文其余部分安排如下：第二章简要回顾相关工作；第三章详细描述SAM的应用范式及提出的MeSAM用于遥感图像语义分割；第四章提供全面的定量和定性实验结果与分析；最后，第五章总结全文。

在缺少提示时，SAM 无法自动生成语义掩码。因此，SAM 提供了一种称为“Everything”模式的自动分割模式，用以为图像中的所有物体生成掩码。然而，“Everything”模式生成的掩码往往缺乏语义信息，导致同一类别的物体被错误地分割为不同的掩码。此外，如图 1(d) 所示，这些掩码可能包含大量过于复杂且无意义的物体掩码。为应对上述挑战，我们提出了一种参数高效微调（PEFT）方法，将多尺度感知适配器无缝集成至 SAM 图像编码器的 Transformer 模块中。尽管 SAM 采用的 ViT 主干网络能够建模整个图像范围内不同补丁之间的长距离关系，但其仅限于同尺度区域之间的交互，缺少跨不同尺度区域捕捉关系的多尺度感知能力，如图 2(a) 所示。与以往通过降低特征空间分辨率并提升通道容量来获取多尺度信息的 Transformer 结构不同，我们提出的多尺度适配器（MSA）能够在整个网络中保持固定的通道容量和分辨率，避免通过降低特征空间尺寸而造成的纹理信息损失，并实现与不同尺度区域特征的交互，使其能够感知不同尺度的信息。该过程在 Transformer 网络内部构建了一个多尺度特征激活金字塔，有效融合了 Transformer 原理与多尺度特征层级的思想，如图 2(b) 所示。具体而言，我们引入了两个多尺度感知适配器，即金字塔多尺度适配器（PMA）和窗口多尺度适配器（WMA），以使模型能够有效捕捉多尺度及领域特定信息。PMA 嵌入于 SAM 图像编码器的全局注意力层，采用金字塔多尺度注意力机制，该机制通过采用不同池化比率的池化操作适配 Transformer 模块，获取多尺度信息。另一方面，WMA 构建于更深层的窗口注意力层，利用串行和并行窗口注意力操作捕捉图像中的内在结构模式，并探索采用不同窗口大小以提取多样且细粒度的多尺度信息。较小窗口尺寸强调局部细节和纹理信息，而较大窗口尺寸则提取更多全局及抽象特征，两者对分割任务均至关重要。为了赋予模型语义感知能力，我们提出优化 SAM 掩码解码器的部分参数，以学习带有类别特定知识的掩码标记，用于指导分割。我们的贡献如下：1）提出了一种用于遥感图像分割任务的新型参数高效微调方法，赋予 SAM 模型感知多尺度物体及各分割物体类别语义的能力；2）引入了两个嵌入于 SAM 图像编码器两种注意力机制中的新型适配器，分别为 PMA 和 WMA，这些适配器提取具备多尺度感知能力及领域特定信息的多尺度特征层级；3）在五个遥感数据集上进行的大量实验证明，我们的方法在利用显著减少的可训练参数的同时，实现了遥感图像分割的新基准。

然而，在遥感领域，SAM在某些应用中表现出良好的潜力，如二元分割和变化检测。具体而言，Sultan等提出了GeoSAM，通过自动生成的多模态提示对SAM进行微调，实现了移动基础设施的分割。张等人引入了UV-SAM，该方法利用一个小型语义分割模型生成用于城中村分割的混合提示，借助SAM试图提升遥感图像分割效果。迄今为止，利用SAM进行语义分割的研究尚属有限。Lucas等提出了一种方法，利用其他多模态大型模型生成适当的提示给SAM，从而实现语义分割结果的生成。类似地，张等采用了相似策略，旨在解决向SAM分割输出中注入语义信息的难题。然而，这些方法高度依赖多模态大型模型解读图像中语义内容的能力，因此在应用于遥感图像时，常因遥感图像在纹理、尺度和场景构成上与自然图像明显不同而表现不佳。此外，此类方法在实际应用中往往伴随较高的计算成本。相比之下，我们提出了一种双分支架构，该架构仅需通过语义分支提供足够全面的提示集，即可更高效且简便地完成以往依赖多模态大型模型的任务。与自然图像相比，遥感图像在语义分割任务中面临更为严峻的挑战。当前主流的语义分割模型通常采用多次下采样操作以提取图像特征，导致遥感图像中小目标的细节信息丢失。此外，现有模型未能充分挖掘和利用遥感图像中的多尺度信息。为了解决上述问题，我们提出了一种创新的语义分割双分支架构，如图1所示。该模型包括语义分支和掩码分支。在语义分支中，我们通过整合卷积神经网络（CNN）和Transformer的混合架构，提高模型对小物体的敏感性。此外，我们引入了一个特征引导（FG）模块，旨在提升模型的能力

使用不同的编码器配合一个通用解码器来处理不同的分割场景。然而，这些方法面临两个主要挑战：1）它们依赖于任务特定的编码器，冗余的结构导致参数开销较大；2）这种静态共享网络无法自适应地适应多场景任务，需要预先了解任务类型以选择合适的编码器。近年来，地理空间基础模型（GFM）作为强大的工具，通过在大规模多样化的地球观测（EO）数据集上训练，已成为学习通用地理空间表示的重要手段[23]。作为通用特征提取器，GFM使得土地覆盖制图和目标检测等任务具备广泛的迁移能力。

由于 SAM 生成的掩码缺乏语义标签，以及自然图像和遥感图像之间的内在差异，SAM 在遥感语义分割中的表现尚不理想 [20]。为解决 SAM 在遥感图像上的局限性，研究人员近期探索了多种方法以提升 SAM 在遥感图像任务中的性能，这些方法大致可分为两类。 1558-0644 © 2025 IEEE。版权所有，包含文本和数据挖掘权利，以及用于人工智能和类似技术的训练权利。允许个人使用，但需获得 IEEE 许可方可重新出版/再分发。详情见 https://www.ieee.org/publications/rights/index.html 。 1) 通过默认 SAM 生成输出：直接使用 SAM 输出作为下游任务的分割结果 [21]，利用 SAM 输出增强分割结果 [22]，[23]，或者将 SAM 输出与少样本学习或零样本学习策略结合使用 [24]，[25]，[26]。例如，[22] 利用 SAM 从高分辨率卫星影像中自动划定农田单元的能力，无需额外训练，大幅提升了美国主要农业区的原始 CDL 数据质量。RS-CLIP [25] 采用包含多个阶段模型微调的课程学习策略，以增强零样本遥感场景分类的性能。 2) 应用提示学习技术以使 SAM 适应遥感领域，从而生成更准确的输出 [17]，[18]，[27]，[28]，[29]。例如，RSPrompter [17] 侧重于学习如何为 SAM 生成提示输入，使其能够自动获得实例级语义掩码，从而增强 SAM 框架的能力。尽管这些进展极大地扩展了 SAM 在遥感领域的应用，但它们通常需要人工设计复杂的微调机制或提示学习策略，以及重新训练遥感语义分割模型或视觉基础模型。为应对上述挑战，我们开发了一个简单而有效的框架，用于利用 SAM 的原始输出对语义分割结果进行后处理。我们的观察表明，尽管 SAM 由于遥感图像与自然图像的差异，可能无法准确分割遥感图像的每一部分，但两者之间仍存在一定的互操作性。通过保留 SAM 表现良好的分割区域，我们可以利用 SAM 从自然图像中学到的有用信息来提升遥感图像分割效果。因此，我们提出了一个简洁的框架，能够准确保留 SAM 对遥感图像边界的识别。结合语义分割模型生成的语义信息，我们能够获得兼具精确边界和语义信息的高质量掩码。我们的方法直接利用 SAM 的原始输出来增强语义分割模型的预测，切实发挥了 SAM 在遥感领域的潜力。本文的主要贡献总结如下。

这些方法包括直接使用 SAM、调整或微调 SAM 以及利用提示学习技术。观察到这些方法主要聚焦于二分类任务，包括二分类分割 [33]，[36]，[37]，[38]，[39]，[40]，[41]，[42] 和变化检测 [43]，[44]，[45]，因为 SAM 生成的掩膜不包含语义信息，无法直接应用于多类别的语义分割任务。另一方面，这些方法通常需要人工设计复杂的微调机制 [42]，[47] 或提示学习策略 [34]，[36]，[44]。因此，它们需要修改并添加由 CNN 或 Transformer 组成的模块，或者针对不同数据集设计特定的提示和微调策略。不幸的是，这些因素增加了 SAM 适应遥感语义分割的难度，需额外的模块或微调。为应对上述挑战，我们从物体和边界两个角度开发了一个简单而有效的框架来利用 SAM。我们的观察显示，由于遥感图像与自然图像之间的差异，SAM 在准确分割遥感图像方面存在挑战。然而，SAM 在识别物体方面表现出较高的能力，如图 1(c) 所示。这些物体实际上是自然地物，或自然地物的一部分。表明 SAM 生成的物体（SGO）能够为识别自然地物提供详尽的物体信息。为了充分发挥其潜力，同时尽量减少对通用语义分割模型的改动，我们提出了一种新颖的损失函数，即物体一致性损失，辅助模型训练。此外，我们引入了边界保持损失 [48]，以促使语义分割模型基于详细的 SAM 生成边界（SGB）信息考虑更优的分割边界。由于这两个损失函数均以通用语义分割模型的输出作为输入，因此无需在解码器末端添加额外的分割头。通过简单且直接地利用 SGO 和 SGB

## 4.2 基于辅助编码与特征混合解码的语义分割方法

### 4.2.1 整体结构与评价模型

说说编码器与解码器的架构。语义分割的评价指标

### 4.2.2 基于卷积神经网络的辅助编码

### 4.2.3 像素解码器特征嵌入

### 4.2.4 多尺度分类预测头

## 4.3 实验结果与分析

本节将实现提出的xxx

### 4.3.1实验参数设置

### 4.3.2实验结果分析

### 4.3.3消融实验

## 4.4 本章小结

xxx具有重要的应用价值。

# .基于SAM提示微调的遥感图像交互式分割

什么是辅助编码与混合解码 怎么作用于语义分割。

## 5.1 引言

## 5.2 基于提示微调与多模态模型的交互式分割

### 5.2.1 多模态结构与交互模型

说说编码器与解码器的架构。语义分割的评价指标

### 5.2.2 多模态融合编码

### 5.2.3 提示微调方法

### 5.2.4 实时交互分割系统

## 5.3 实验结果与分析

本节将实现提出的xxx

### 5.3.1实验参数设置

### 5.3.2实验结果分析

### 5.3.3对比实验

## 5.4 本章小结

xxx具有重要的应用价值。

# 总结与展望

## 5.1 本文总结

在当前的信息时代，人工智能技术的迅猛发展以及边缘设备的普及，xxxx，也为进一步研究提供了新的思路和方向。

## 5.2 未来工作展望

本文针对xxxx进行了研究，并提出了相应的解决方案。然而，这些方法仍然存在一些不足之处。为提升本文方法的安全性、实用性和泛化能力，作者计划在未来开展以下研究：

1. xxx。

2. xxx。

3. xxx。

4. xxx。

综上所述，xxxx，为未来的智能社会构建更安全、高效和智能的基础设施。

# 致 谢

。

# 参考文献

Dinh C T, Tran N H, Nguyen M N H, et al. Federated learning over wireless networks: Convergence analysis and resource allocation[J]. IEEE/ACM Transactions on Networking, 2020, 29(1): 398-409.

# 附 录

**攻读硕士学位期间发表的论文和出版著作情况：**

专利

**论文**

**攻读硕士学位期间参加的科学研究项目情况：**

Xxxx

**攻读硕士学位期间竞赛获奖情况：**

[1]